```r
# load packages -----------------------------------------------------------

## Package names
packages <- c("tidyverse", "here", "papaja", "hablar")

## Install packages not yet installed
installed_packages <- packages %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages])
}

## Packages loading
invisible(lapply(packages, library, character.only = TRUE))
```

```
## -- Attaching packages ----- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts -------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## here() starts at C:/Users/keana/OneDrive - PennO365/Comp_transfer2018/Penn/stats_masters/stats-master

##
## Attaching package: 'hablar'

## The following object is masked from 'package:dplyr':
##
##     na_if
```

```r
filter <- dplyr::filter
```

Participants

The final sample consisted of 527 participants (290 women) from Amazon Mechanical Turk. We did not exclude any participants across the analyses. Ages ranged from 19 to 82 years, with an average age of 39.9013283 years (*SD* = 13.2160224). We included participants based upon the following criteria: (a) adults on (b) Amazon Mechanical Turk) (c) born and currently residing in the US (d) have had 90% or greater of their previous HITs approved, and (e) have a device with audio capabilities. We excluded Black individuals during the pre-screening process, since we are primarily interested in understanding the factors that affect threat and leadership perceptions of Black men, and group membership may differentially affect these perceptions.

Design

The study was a 2X2 within-subjects design with two independent variables: voice pitch (high or low) and race (White or Black names). Each of the four conditions was counterbalanced. Names and individual voices were randomly assigned to participants without repeat. This ensured that individuals would not listen to a high and low voice that resulted from the same original voice.

Procedure

Participants were recruited through a HIT (human intelligence task) posted on Amazon Mechanical Turk. They were told that they would listen to a participant that previously provided their recording and took a "series of character trait and performance tests," which would then be compared to the participants' ratings to assess the accuracy of their perceptions. Upon being assigned to a recording, they learned the participant's name, and were provided other information about the recording (i.e., location, date) to make the design less conspicuous. All of the names were randomly assigned to correspond to the high-pitched or the low-pitched conditions. The presentation of the four names for the recordings was randomized and counterbalanced across participants. Then, they listened to the participants' recording by clicking on the Soundcloud file embedded in the survey.

They were asked to assess the participant's character based upon their voice using a series of 100-point slider scale questions (i.e., trustworthiness, dominance, threateningness), which served as our measures of perceived trustworthiness, perceived dominance, and perceived threat, respectively. The presentation of the scale items was counterbalanced for each participant and within each condition. Additionally, we asked them to rate the individuals in the recording on various traits that were independently rated as important for leaders on 100-point slider scale items. Finally, they indicated their preferences for engaging in different types of interdependent relationships with the people in the recording on 100-point slider scales. Participants could listen to the recordings as frequently as they desired before rating the voices. They completed demographic questions and indicated what they thought the study was about as a suspicion check. After participants completed the suspicion check, we determined whether the manipulation of the names elicited perceptions of the race of the recorded individuals through a series of manipulation check questions. First, we created a name attention check score based upon whether the participants remembered the names of the people in the recordings. The participants were presented with a list of eight names, four of which were included in the study. Every time they correctly identified a name that was presented to them during the study, they received a point, for a total name recall score of four points (M = 3.1157495, SD = 0.9768609). If they incorrectly selected a name that was not presented to them, they did not receive a point. Participants were asked how many people in the recordings they thought were White (M = 2.91, SD = 0.86, with 13.28% of participants indicating they were unsure) or Black (M = 0.99, SD = 0.82, with 15.56% of participants indicating they were unsure). Finally, we asked participants to rate the likelihood that people with the names used in the study would be White or Black on 100-point scales (see Figures @ref(fig:f3) and @ref(fig:f4)). A debriefing page explaining the true purposes of the study and the logic behind the deception was provided before payment. Participants were paid $1.00 for their participation.

Materials and measures

Voice stimuli

For the voice stimuli, we recorded the voices of eight White men between 18-30 years of age in Audacity using the Zoom H4N Handy Recorder with a sampling rate of 44.1kHz. The men quoted the first sentence of the Rainbow Passage (e.g., "When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow") [@Fairbanks1960]. At the end of each sentence, the men read a randomly assigned identification number provided by the researchers. The four-digit identification numbers were created randomly, and participants were required to enter the identification number as a means of verifying that they were listening to the recordings.

After the recording sessions, each voice was manipulated to have a higher or lower pitch in Praat (Version 6.0.36) [@Boersma2001], which served as our manipulation of threat potential through the voice. We followed the standard methods in voice research by raising and lowering each voice by 0.5 equivalent rectangular bandwidths (ERBs) using the Pitch-Synchronous Overlap Add tool in Praat, which produces a shift in perceived pitch of approximately 20 Hz in either direction [e.g., @Apicella2009; @Klofstad2012; @Tigue2012; @Vukovic2010]. We set the pitch floor to 70 Hz and the pitch ceiling to 250 Hz, which has been validated as an appropriate range for male voices [@Vogel2009]. Many researchers manipulate ERB instead of Hertz because a change in pitch is perceived differently depending upon the original pitch that was manipulated, since there is a logarithmic relationship between actual pitch and perceived pitch (Stevens, 1998). Also, the ERB manipulations will not affect other acoustic characteristics of the recording (e.g., rate, intensity) [@Feinberg2005]. Since each of the voices was raised and lowered in pitch, there were a total of sixteen manipulated recordings. We checked the manipulation by comparing the mean pitch for the original voices ($M = 104.37$, $SD = 14.09$) to the lower manipulations ($M = 90.21$, $SD = 9.79$) and the higher manipulations ($M = 121.30$, $SD = 17.28$). All of the manipulated files were uploaded to separate Soundcloud links and embedded in the survey.

Names for race manipulation

To manipulate perceptions of race, we used four names that are typically associated with Black people (i.e., Tyrone, Keyshawn, Deshawn, Terrell) and four names that are typically associated with White people (i.e., Scott, Brad, Brett, and Logan) [@Gaddis2017]. Names were presented before the participants listened to the voice recording. Each name was chosen based upon the criteria that 90% or more of raters from @Gaddis2017 thought that the individual was either Black or White when they were asked about their perceptions of the person's race based upon their name.

Perceived leadership ability

We recruited 55 participants on Amazon Mechanical Turk to serve as independent raters for identifying the leadership traits used in the experiment. We provided them with a list of fifteen traits from which they could select what they considered most valuable for successful leaders of businesses and companies (e.g., drive, creativity, confidence) [@Kirkpatrick1991]. We selected the traits that were ranked, on average, in the 30th percentile of responses (where 1 is considered the most important trait for a leader). The traits that were selected for the leadership composite score based upon these criteria were intelligence, effective communication, confidence, and problem-solving ability, which were rated by participants in the final study using 100-point slider scale items.

To create the leadership ability composite, we averaged participants' ratings of the individual in the recording on the four traits. Higher scores denote greater perceived leadership ability. The measure had high internal consistency across participants in the final sample ($\alpha = 0.9105731$; average alpha across conditions).

Perceived threat, trustworthiness, and dominance

Single questions were used to elicit perceived threat, trustworthiness, and dominance. Participants responded using a 100-point slider scale.

Multilevel models

Given the hierarchical nature of the data (e.g., condition nested within subjects), we employed multilevel models (also known as linear mixed-effects models or hierarchical linear models) [@Finch2014; @Raudenbush2002; @Gelman2007; @Garson2013; @Galecki2013] to analyze the data. The basic premise of using this type of analysis is to account for the inherent correlation among the observations nested within other variables. For instance, within the context of the current study, we measured participants' rating of threat across all conditions, so it is likely perceptions of threat within each participant will be correlated, since there may be inherent individual differences in participants' baseline perceptions of others' threat and/or perceptions of threat in response to each combination of race and voice pitch. If we did not account for this correlation in responses within each participant, we would be violating the assumption of independence of observations. Although repeated measures analysis of variance (ANOVA) is often used for analyzing data of this nature, we use multilevel models in the current paper because they present several notable advantages over repeated measures ANOVA. For instance, multilevel models are more powerful in the face of "unbalanced" repeats, where the measure of interest is missing one or more observations. Repeated measures ANOVA employs listwise deletion in the face of missing data points, reducing the effective sample size. On the other hand, multilevel models use the data available within a group to estimate parameters and compute inferential statistics, while accounting for the fact that some estimates are more reliable than others [@Raudenbush2002; @Brauer2018; @Misangyi2006]. Notably, we assume that the data points are missing at random for these inferences. Additionally, multilevel models allow researchers to explore multiple groups with correlated observations (that is, multiple sources of nonindependence) [@Brauer2018; @Westfall2014; @Westfall2015], while repeated measures ANOVAs only allow one to account for one source of nonindependence [e.g., @Baayen2008; @Judd2017]. The effects of continuous predictors that may vary within groups can only be analyzed using multilevel models, since repeated measures ANOVAs only accept categorical predictors [@Misangyi2006; @Brauer2018]. Finally, multilevel models allow researchers to explicitly model different sources of variation within the data (e.g., individual and group-level variation in group-level estimates, variation in individual-level estimates) and estimate the effects for specific groups [@Gelman2007].

Model estimation and comparison

Multilevel models are a variation of classical regression that assign a probability model to specific regression coefficients [@Gelman2007]. The parameters of this second-level probability model have their own coefficients, known as hyperparameters [@Gelman2007]. Although classical regression has the capacity to model varying coefficients with the use of indicator variables, multilevel models are unique in their ability to model the variation between groups by including varying coefficients and models for each varying coefficient [@Gelman2007]. To model variation at multiple levels, these models incorporate what are known as "fixed" and "random" effects (although see @Gelman2007 for a discussion on the various names used for these effects), where random effects are typically conceptualized as effects that vary across the nested groups, while fixed effects are constant across all groups within the data [@Finch2014].

When analyzing data using multilevel modeling, there are different random effects structures that can be used to model the data. A random effects structure is essentially the way the parameters are assumed to vary across the nested groups [@Barr2013]. The most basic random effects structure includes only a random intercept, which allows the intercept to vary across groups (i.e., there is a different intercept estimated for each group). A more complex random effects structure allows the slopes to vary by group (i.e., fitting a unique regression line to each group) in combination with the random intercepts. There are many different ways to model the random effects structure (e.g., random slope by group with correlated intercepts, random slope by group without varying intercepts by group, uncorrelated random intercept and slopes by group, etc.) [@Meteyard2020], which will change the interpretation of the results and may even reduce power or increase Type 1 error [@Barr2013; @Matuschek2017], so it is important to identify a random effects structure that is appropriate for the data.

The most appropriate way to determine the random effects structure for one's data is still debated. Some researchers have argued that it is imperative to fit "maximal models" (i.e., fit random slopes, including interactions, and intercepts for each predictor in the model) whenever possible, and only reduce the random effects structure when the model does not converge [@Barr2013]. When a model does not converge, it essentially means that the optimization algorithm used to estimate parameters cannot reliably determine the maximum likelihood function for that model [@Brauer2018]. This typically occurs when there is insufficient data for the number of parameters being estimated. Thus, failures to converge are much more likely to occur when trying to estimate maximal models with many terms. On the other hand, fitting models with only intercepts varying by group leads to inflated Type 1 error [@Schielzeth2009; @Barr2013]. Thus, @Barr2013 argue that the common practice of fitting varying-intercept only models can lead to biased conclusions and recommend fitting the maximal model that will converge. However, it has been argued that the rise in employing maximal models can lead to their own set of problems, namely i) failure to converge [@Bates2015], ii) models that converge but are so overparameterized that they are uninterpretable [@Bates2015], and iii) loss of power due to random effects contributing little to the model [@Matuschek2017]. In place of "maximal" models, @Matuschek2017 and @Bates2015 argue that one should employ "parsimonious mixed models" when analyzing their data, where the researcher uses a pre-determined model comparison technique (e.g., likelihood ratio test, Akaike information criterion, Bayes/Schwarz information criterion) to select the random effects structure that best fits the data. To employ the parsimonious mixed model approach, one would first fit a maximal model, then remove random effects that are not contributing to the model (i.e., variance is close to 0), stopping before they reach a model that would significantly reduce the goodness of fit [@Matuschek2017; @Bates2015]. In support of this argument, @Matuschek2017 use simulations to demonstrate that parsimonious models can reduce Type 1 error associated with underfitting models, while attaining higher power than maximal models. This is because maximal models can lead to a decrease in statistical power if they have random effects that do not contribute to the fit of the model (e.g., with variances near 0) but reduce the degrees of freedom, essentially increasing the standard errors of the fixed effects estimates [@Matuschek2017]. At the same time, it is generally accepted that random effects with near-zero variance do not affect goodness of fit tests [@Brauer2018]. Despite the lack of consensus regarding how to determine the final model, most researchers suggest starting with the maximal model and that the final model should have an effects structure that aligns with the researcher's main hypothesis, even if these random effects have near-zero variance [@Brauer2018; @Barr2013; @Bates2015].

Another general point of consensus is that one should use restricted estimated likelihood (REML) instead of maximum likelihood (ML) estimation for unbiased estimates of the random effects parameters [@Brauer2018; @Maas2005; @Browne2006; @Elff2020; @Gelman2007], especially with smaller samples at the group-level [@Hox2020;@Mcneish2016a;@McNeish2017]. The problem with ML estimation typically arises with smaller samples because the process of ML estimation tends to ignore variability in the fixed effect estimates and does not account for the degrees of freedom used to estimate the fixed effects [@McNeish2017]. Thus, ML estimation can lead to more bias in random effects estimation with smaller samples because the effects in these cases tend to be more sensitive to small changes in the degrees of freedom and tend to have larger sampling variability. Since random effects parameters are estimated based on the fixed effects parameters, this can cause the random effects to be underestimated [@McNeish2017]. As a result, the standard errors of the fixed effects tend to be underestimated because the random effects estimates are integrated into the formula for fixed effects standard errors. With smaller standard errors, the t or Z test statistic will be overestimated, leading to higher Type 1 error. The process employed by REML estimation leads to better estimates of the random effects, which in turn improves the fixed effects standard error estimates [@McNeish2017].

Finally, researchers have examined how different techniques for evaluating significance of effects in multilevel models affect Type 1 error rates. Notably, @Luke2017 show through simulations that likelihood ratio tests (LRTs) and applying the Z distribution to the Wald t values from the model output can lead to unusually high Type 1 error rates, especially with smaller samples (i.e., less than 40-50 number of items and/or subjects). This is the case when fitting models using both ML and REML estimation. Of the options available to researchers in the statistical software *R*, Type 1 error is closest to .05 when deriving p-values using Kenward-Roger [@Kenward1997] or Satterthwaite [@Satterthwaite1941] corrections for approximating denominator degrees of freedom for F statistics or degrees of freedom for t statistics [@Luke2017]. Although these corrections tend to produce similar output [@Luke2017; @Elff2020], @Hox2020 and @McNeish2017 argue that the Kenward-Roger provides slightly better approximations by correcting standard error and estimating degrees of freedom, while the Satterthwaite correction only estimates the effective degrees of freedom. However, after comparing results from each method, @Elff2020 recommends the Satterthwaite method because it produces results that are indistinguishable from the results produced by the Kenward-Roger method, with far less of a computational burden.

Sample size considerations

Another important consideration in determining model structure of multilevel models is the sample size at each level (i.e., number of groups and number of individuals within each group). Like in most parametric statistical inference, the estimates become unreliable, or even impossible to estimate (if the model fails to converge), with sparse data. These effects may differ depending on the level, where @Scherbaum2009 showed that increasing level-2 sample size (i.e., number of groups) had a larger effect on variance components than increasing level-1 sample size (i.e., number of individuals within each group).

Understanding how “sparse” one’s data can be at each level while being able to maintain unbiased estimates has been the subject of several lines of recent work. One of the seminal pieces in this literature suggested that level-2 standard errors are biased when the sample size is less than 50 (i.e., there are fewer than 50 groups) [@Maas2005]. Other recommended standards are to have 10 observations for at least 100 groups to estimate a random intercept for said group, and at least 20 observations with a minimum of 200 groups for estimating slope variance [@Clarke2007]. @Scherbaum2009 recommends 30-50 trials/items per participant for power. Notably, the size of each group can affect estimation of the random effects, but tend to have little to no impact on estimation of fixed effects [@Clarke2007; @Maas2004; @Maas2005].

Thus, recent work has focused primarily on the effects of small samples on the estimation of random effects. As of recently, many researchers are suggesting that, under certain conditions (e.g., continuous outcome variables, five or fewer fixed effects, no missing data, one or two variance components), there can be as few as 7-10 groups at the second level to be able to estimate random effects with reasonable accuracy using REML estimation with cross-sectional data. However, the appropriate sample size will intrinsically depend on the nature of the data and model at hand [@Hox2020].

With small samples, some suggest that Bayesian estimation can be a more accurate alternative [@Stegmueller2013] because Bayesian statistics do not rely on the central limit theorem [@Hox2020]. Specifically, @Stegmueller2013 performed a Monte Carlo experiment to compare the performance of frequentist and Bayesian multilevel models when there are few (e.g., 5, 10, 15) groups, and showed that the frequentist approach tended to be anti-conservative and biased with smaller samples. However, @Elff2020 has recently argued against the notion that standard multilevel models are inferior to models following a Bayesian framework with a small number of groups. They showed that the estimation bias found in @Stegmueller2013 could be solved by using i) REML estimators for variance parameters and ii) a t-distribution with appropriate degrees of freedom for statistical inference. Through these relatively simple steps, the standard multilevel models were found to produce unbiased estimates of both fixed and random effects.

Additionally, any possible advantages of Bayesian estimation are completely dependent on the choice in a prior probability distribution (often simply called a prior) [@Gelman2007; @Hox2020]. A prior essentially represents the knowledge, a priori, one has (if any) about the distribution of the parameters, which is then combined with the data observed to produce posterior inferences. Thus, the choice of a prior probability distribution is critical, especially with smaller samples [@Mcneish2016]. For instance, in a systematic review of the literature on estimation while using small samples, @Smid2020 showed that inference with uninformative priors can lead to estimates that are just as, if not more, biased than the estimates from frequentist methods when working with a small number of groups. Unfortunately, the default prior distribution in most software is uninformative, so it is entirely possible that many researchers acquire biased estimates by using Bayesian estimation naively [@Hox2020].

Multilevel modeling methodology for current research

We used R [Version 3.6.3; @R-base] statistical software for all analyses (see here for all data and analysis scripts and see Appendix for full list of R packages used). To conduct analyses for the current work, we employed many of the techniques and recommendations described in the above sections. For instance, in determining the maximal random effects structure, we followed the guidance of @Brauer2018 and @Barr2013, who argue that i) every source of nonindependence should be modeled through a random intercept ii) generally, there should be a random slope for each within-unit predictor and iii) one should estimate random slopes for interaction effects when all factors comprising the interaction are within-group. They also note that there are exceptions to these general rules of thumb. For instance, i) does not need to be followed if the purported random effect is fully confounded with a predictor in the model. That is, when the random variable is nested within a fixed effect, we do not need a random intercept for this variable. Since the names chosen for the current experiment were necessarily nested within race, we do not model random effects for name. Also, these general guidelines are under the assumption that there are enough data at each level of the model to be able to obtain reasonable estimates. In our case, we have 527 participants, which means that we are likely well-powered to estimate random effects for each participant (that is, we have 527 groups for the participant variable), even with four observations within each group. As mentioned before, the level-2 sample size is more important in determining the power than level-1 sample size [@Scherbaum2009]. However, we also assume that the voice presented to participants in the recording is a random effect, since it is likely that responses to one voice will be more similar to each other than responses to another voice. One point of concern with obtaining estimates for the voice variable is that the sample is relatively small, with only 8 voices used during the experiment. Since we attempt to estimate random slopes and intercepts for the voice variable and may be underpowered (especially for random slopes) [@Hox2020], we recommend taking the sample size into consideration when interpreting the estimates for the random effect of voice and encourage future research with larger samples to determine whether these results are replicable. Across all models, we recoded the race and voice pitch variables to improve interpretability of the coefficients [@Brauer2018]. For the race variable, Black names were recoded to -.5 and White names were recoded to .5. Within the voice pitch variable, -.5 corresponded to a low-pitched voice, while high-pitched voices were recoded as .5.

Here, we will describe the steps for selecting our final random effects structure. For each model, we started with a maximal random effects structure, as defined above. Although it is generally recommended to fit a model with a random effects interaction term for any within-subjects fixed effects that have an interaction term [@Barr2013], in this case, we did not have enough data for fitting the interaction between two random within-subjects slopes (i.e., voice pitch and race), because there would have been one data point per cell, which confounds variability between the conditions with variability between groups [@Brauer2018]. Instead, we start with maximal models that have all within-subjects slopes and intercepts possible. Then, we submitted this maximal model to the buildmer() function from the buildmer package [@R-buildmer] with a bound optimization by quadratic approximation (BOBYQA) optimizer to select a final model. Using the maximal model formula provided, the buildmer function enters parameters one-by-one in terms of their contribution to the change in log-likelihood (i.e., terms that have lower chi-square p-values are entered first), stopping when a model does not converge. By following these steps, the function ensures that the most relevant parameters are included in the model before the model fails to converge. After selecting this maximal model, the function follows a backwards step-wise selection process, where terms are eliminated when they do not significantly contribute to the model (that is, when removing the parameter does not cause a significant reduction in the likelihood-ratio). Once the final models were selected, we performed assumption checks for each model (e.g., normal distribution of residuals, homogeneity of variance), which are described in the Appendix for all final models and should be also be considered when interpreting the effects.

For each model, we also tested the random effects assumption - where we assume that the random effects within the model are not correlated with the fixed effects. The random effects assumption is an essential, yet underappreciated assumption when running multilevel models [@Antonakis2019]. If this assumption is violated, it can introduce endogeneity into the model, which ultimately can lead to biased and inconsistent estimates. To test this assumption, we ran a Hausman's test [@Hausman1978] with a fixed effects estimator and a random effects estimator using the phtest() function within the plm package [@R-plm]. The Hausman test will assess whether model estimates were consistent with fixed effects, which include a dummy variable for groups as a predictor in the fixed part of the model, or random effects models, which include random intercepts for each group. If the p-value of the test was significant, this would suggest that the random effects assumption was not met and that we would need to switch to a correlated random effects model (CRE), which includes the cluster-means for each grouping variable as a control variable and, in doing so, makes the random effects conditionally independent from the fixed effects within the model [@Antonakis2019]. If the p-value was not significant, we proceeded with the original random effects model. We report random effects models for all analyses unless otherwise specified.

P-values and degrees of freedom in each multilevel model were calculated based on Satterthwaite's correction using the lmer() function within the lmerTest package [@R-lmerTest] and BOBYQA as the default optimizer. We also ran robust versions of the models as a point of comparison using the rlmer() function from the robustlmm package [@R-robustlmm] (see Appendix for comparison of maximal, final, and robust models). The robust models account for multiple possible sources of "contamination," wherein outliers at the higher levels of the model can produce contamination at the individual level, by robustifying the estimating equation through huberization and applying the Design Adaptive Scale approach [@Koller2016; @Koller2011] to get robust estimating equations of θ and σ.

In the subsequent results section, the fixed effects within each model will be the focus of the analyses, so I encourage readers to refer to the respective tables within the Appendix if interested in the random effects. Here, we will briefly summarize the notation for the random effects summary tables provided in the Appendix to aid interpretation. The first statistic provided in the random effects parts of each table (σ^2) is the residual variance that is not explained by either the random or fixed effects within the model (that is, the general ϵ in most linear models). Then, the variance attributed to each portion of the random effects within the model, including the slopes and intercepts for each group, is listed. For instance, in Figure @ref(fig:f5), the value of τ_{00id} (i.e., 128.30) reflects the variability of the intercept across participants. Similarly, $\tau_{00id.cond\_raceC}$ reflects the amount of variability in the slope for the race variable across participants.

Next, if a correlated random slope and intercept is included in the model, the table presents the correlation between the intercept and the associated random slope. For instance, $\rho_{01id.cond\_raceC}$ in Figure @ref(fig:f5) shows that there is a negative correlation between the random intercept for participant and the random slope for the race variable. That is, for each one-unit increase in standard deviation of a participants' intercept, the model suggests that the participants' slope would decrease by 0.13 standard deviations.

Afterwards, the tables list the intra-class correlation coefficient (ICC), which reflects the proportion of the variability in the dependent variable that can exclusively be attributed to the effects of the groups. For instance, ICC in Figure @ref(fig:f5) suggests that 43.82% of the variation in perceptions of threat can be predicted solely based on the participant that was giving the ratings and the voice that was listened to.

Then, the tables list the size of each group. For instance, the effective sample size for the by-participant random effects is 527, while the by-voice random effects have an effective sample size of 8. As suggested by the final parts of each table, across all dependent variables, there are a total of 2108 observations, since we measured each variable 4 times for each of the 527 participants. Finally, the marginal and conditional R^2 are listed, where the marginal R^2 is the "proportion of the total variance explained by the fixed effects," while conditional R^2 is the "proportional of the total variance explained by both fixed and random effects" [@Nakagawa2017]. Notably, for the tables comparing the maximal models, final, and robust models, several of the maximal models have a conditional R^2 that is listed as "NA" and a correlation between intercepts and slopes (ρ) of +/- 1, which suggests that the model was singular, which can lead to unreliable estimates [@Bates2015]. Thus, we recommend exercising caution in interpreting the effects found in the maximal models with singular fits. We include these models in the table primarily for the the purpose of transparency, to show the model with which buildmer() started to estimate the final models used in the primary analyses.